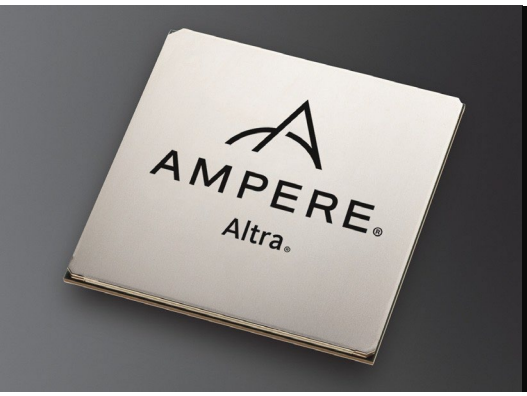

Object Detection with YOLOv8 on GCP Tau T2A VMs

GCP T2A VMs, powered by Ampere® Altra CPU and high performance Ampere® AI inference engine, deliver best-in-class GPU-Free AI inference performance on standard AI frameworks, including PyTorch, TensorFlow, and ONNX-RT.

Ampere Altra Powered ML Inference on GCP

Ampere® Altra family of **Cloud Native Processors** meet the needs of widely used machine learning (ML) workloads while **providing the best price-performance and optimized power draw**. This demo consists of multiple streams of video sources detecting still and moving objects such as apples and water bottles using the popular YOLOv8 model. It also showcases how this implementation can be used for a task of counting the number of units.

Setup

Deployment of the open-source computer vision object detection AI model YOLOv8 with Ampere® Optimized PyTorch running on Ampere Altra Max. The chosen model, YOLOv8, is a widely used algorithm for computer vision applications where both throughput and latency are critical.

Key Benefits Demonstrated

- Meets or exceeds the necessary **low latency** requirements for real-time ML object detection applications.
- Delivers the **best price-performance** in GPU-Free AI inference in both cloud deployment scenarios.
- The YOLOv8 model can be downloaded from Ampere® AI Model Library (AML) and used as is without any modifications.
- Ampere Altra processor can be **easily scaled** and **dynamically provisioned** based on the performance requirements of the user's application such as target frame rate, number of video channels, etc.

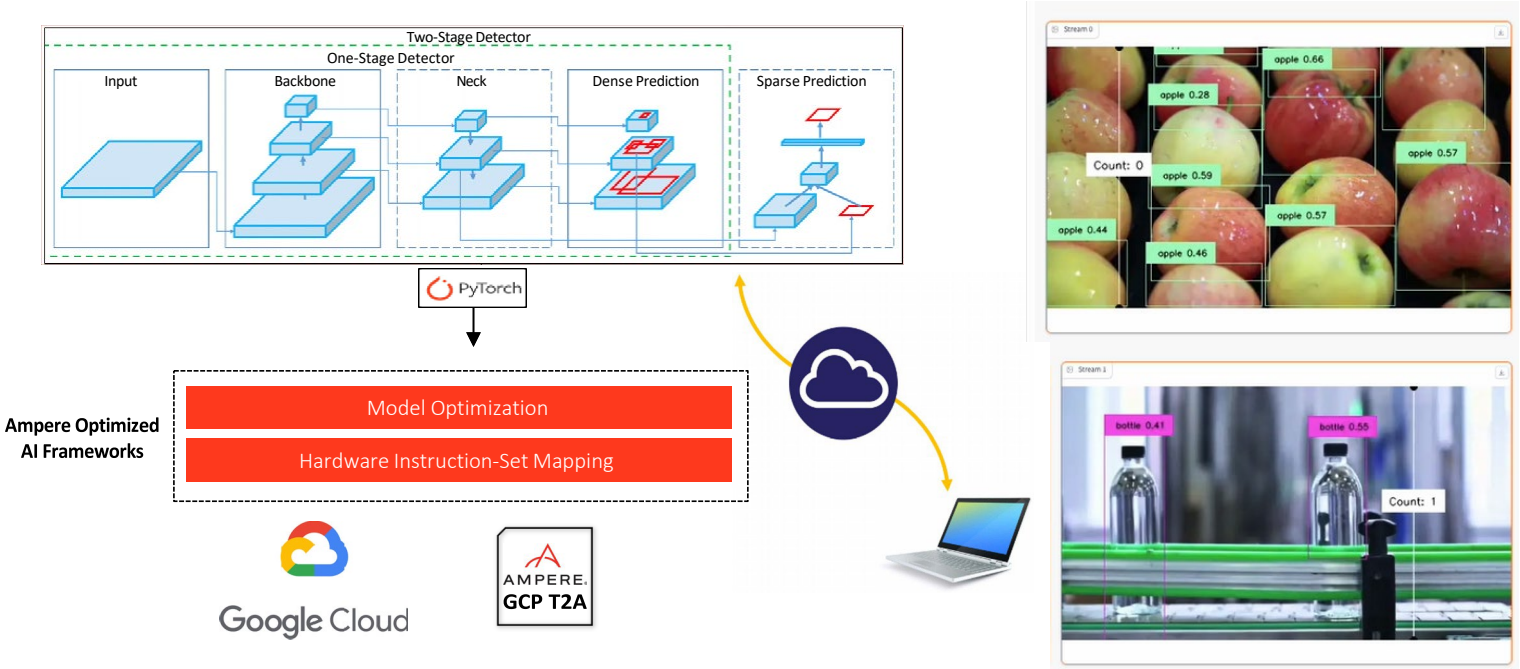

Figure 1: YOLOv8 demo runs on GCP Tau T2A Instance with Ampere Altra

Real-time Object Detection and Classification

This demo performs object detection and classification with a pre-trained YOLOv8 model. It processes images and videos from an incoming real-time video streaming from video files. It runs on a **GCP T2A VM** at **real-time performance level**. The performance can be scaled depending on application requirements by allocating the number of vCPUs to meet the desired price-performance target.

The same workload also runs on x86 for comparison purposes. We demonstrate that the **Ampere Altra family of Cloud Native Processors consistently outperforms x86 platforms.**

Resources

The YOLOv8 model can be accessed from the Ampere AI Model Library. Ampere Optimized Pytorch, and other Ampere Optimized AI Frameworks, can be accessed directly from Google Cloud Marketplace.

The docker images are also available in the downloads section of Ampere AI Solutions web page. All software is available free of charge and runs straight out-of-the-box with no additional coding required.

Figure 2: The integration of Ampere Optimized AI Frameworks with Ampere Altra Cloud Native Processors